Knowl Inf Syst
DOI 10.1007/s10115-016-0970-8

REGULAR PAPER

Community detection in social networks using user frequent pattern mining

Seyed Ahmad Moosavi[1] · **Mehrdad Jalali**[1] · **Negin Misaghian**[2] · **Shahaboddin Shamshirband**[3] · **Mohammad Hossein Anisi**[3]

Received: 28 April 2015 / Revised: 17 February 2016 / Accepted: 27 May 2016

Abstract Recently, social networking sites are offering a rich resource of heterogeneous data. The analysis of such data can lead to the discovery of unknown information and relations in these networks. The detection of communities including 'similar' nodes is a challenging topic in the analysis of social network data, and it has been widely studied in the social networking community in the context of underlying graph structure. Online social networks, in addition to having graph structures, include effective user information within networks. Using this information leads to enhance quality of community discovery. In this study, a method of community discovery is provided. Besides communication among nodes to improve the quality of the discovered communities, content information is used as well. This is a new approach based on frequent patterns and the actions of users on networks, particularly social networking sites where users carry out their preferred activities. The main contributions of proposed method are twofold: First, based on the interests and activities of users on networks, some small communities of similar users are discovered, and then by using social relations, the discovered communities are extended. The *F*-measure is used to evaluate the results of two real-world datasets (Blogcatalog and Flickr), demonstrating that the proposed method principals to improve the community detection quality.

Keywords Social networks · Community detection · Frequent pattern mining · Data mining · Big data analysis

✉ Mehrdad Jalali
Jalali@mshdiau.ac.ir

Shahaboddin Shamshirband
shamshirband@um.edu.my

[1] Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

[2] Young Researchers and Elite Club, Mashhad Branch, Islamic Azad University, Mashhad, Iran

[3] Department of Computer System and Information technology, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

Published online: 27 July 2016

1 Introduction

It has been more than two decades that interactions among artists and determining important structures in communications on social networks have been analyzed [7]. It is possible to view social networks from different aspects, for instance systems such as Facebook, which is explicitly designed for social interactions, or Flickr that is designed to offer various services such as sharing content and widespread social interaction for users. A social network is depicted in a graph (Fig. 1), where the nodes consist of actors, and the edges represent the relationships or interactions between these actors [7,10,48]. It is thus evident that social network content is not limited to networks like Facebook, Flickr, Twitter, etc. The aim of this study is to analyze Internet social networks and discover communities throughout these networks.

Social network Web sites on the Internet offer significant potential of communication and interaction between people who are geographically spread out, linking them in different ways. They also facilitate interaction and sharing of information with different people including relatives, co-workers, family, friends, fans, and others [36]. In addition to facilitating communication, social networking sites allow updating, liking, disliking, creating profiles, and sharing personal and public information [42].

The structure of social networks is a good index for predicting potential attempts of users. Thus, there are different modes of operation in Internet social networks in addition to the interaction among users. Users normally choose the operations according to their taste, which is what establishes complementary network structures.

One of the fundamental challenges regarding analyzing social networks is the automatic discovery of communities [19]. Communities are seen as groups, clusters, subgroups or moduli in various areas, and discovering a community in a social network means recognizing a set of nodes communicating with each other more than other nodes in the network. Simultaneous to the rapid development of the Web in a social sense, Web sites for social networks are designed in new forms that enable people to communicate with others. Community detection on such Web sites can facilitate other social computation duties and application in most programs. As an example, we can refer to group of customers based on similar benefits from social media, effective suggestions of goods, proposal systems that are conventional in social media programs. Community structures can be considered as a summary of the whole network thus easy to visualize and understand. In this work, a method is suggested for discovering communities on Web sites of social networking. This method emphasizes on friendship information among users and individual user information to discover communities.

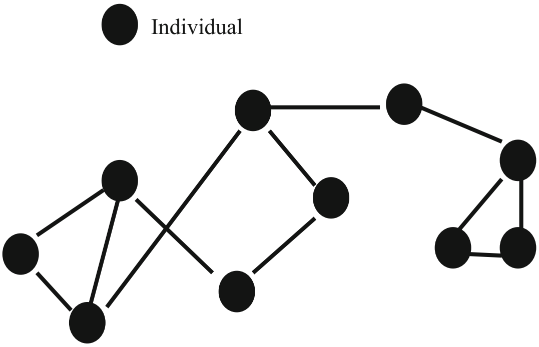


Fig. 1 Sample social network

According to popular social networking sites, the numbers of network users have grown significantly, so that today social networks represent massive databases of user information. The discovery of unknown patterns in such networks can be useful in many applications, and community detection can be employed in various fields. In every area, it is basically possible to create a network of entities that are connected by one or more specific types of dependencies. Such network can be displayed in the form of a graph that is often large and complex. Important and latent information can be attained by identifying dense structures or similar nodes. Detecting communities is useful in various fields of social networking applications such as friend suggestions, customer segmentation, link derivation, tagging vertices, and social influence analysis [41].

A considerable amount of research has been done on solving these challenges and provided different algorithms.

The most well-known methods of dealing with the challenges of community detection only take into account connection information to predict communities [41]. For instance, a community in biology that comprises proteins, genes, or similar sub-units shows community members with similar behaviors and actions.

Consider a community or social networking sites whose members have similar characteristics and more interactions. Detecting community structure is useful for visualizing social networks. Core groups of users and their interactions can serve to display social networks [37].

Recognizing specific structures has many applications in a range of areas, such as routing in networks [13] and releasing worms in cellular networks [54].

The classification of nodes and recognizing leaders or vital connectors in a group are other applications of community detection in social networks. Detecting leaders or influential individuals and their societies may be an effective solution to extend product advertisement in a network.

Finding influential people in a science social network, e-mail network, discussion group, etc., assists with analyzing the networks [28,32]. For instance, a community is World Wide Web pages that are grouped according to their related subjects [17], functional modules such as cycle and paths in metabolic networks [22] or groups of people that are connected in social networks [31]. Most researchers in the field of community social network believe that the detection of communities as a traditional data mining task is comparable to the problem of clustering in data mining.

Clustering in data mining is an unsupervised type of learning, which aims to divide a large dataset into different homogeneous groups (clusters) [31]. In fact, the discovery of a society can be considered data mining on graphs.

In addition, the detection of communities is the largest study area of data mining applications in social networks. Other applications like graph mining are still in early development stages [50].

However, since the aim of this study is to detect communities in social networking Web sites, in addition to existing communications, other rich content in these networks can be used in order to improve the quality of the communities. Besides the relationships between users, there is an abundance of valuable information on social networking Web sites.

Some recent works have shown that the use of edge or node content from social networks can be beneficial to improving the quality of communities. The main objective of this study is to offer a method of detecting communities, where along with data connection among nodes, content from social networking Web sites can be employed.

With this method, the goal is to detect community quality Web sites for use in applications such as friend suggestions and customer segmentation to improve the effectiveness of communities.

If communication were to only consider people who are interconnected can be identified as a community, perhaps congenial and like-minded people on a network, there would be a loose connection. Thus, the use of additional information to detect this structure can be advantageous. As such, the framework presented in this study is a new approach in the field of community detection.

The remaining sections of the present work are organized as follows. Section 2 introduces related works, after which a framework for discovering communities is presented in Sect. 3. The experimental tests are described in Sect. 4 besides how this approach is effective in discovering Internet communities. Advantages and disadvantages of the proposed method are described in Sect. 5, and finally, a discussion and suggestions for future attempts conclude the study in Sect. 6.

2 Related work

The problem of the importance of community discovery in social networks has been widely studied [1,9,18,21,26,27,30,34,35,40,41,44,52,55]. Methods of discovering communities are based on agglomerative clustering, min-cut-based graph partitioning, clique percolation methods, and the measure of social networks analysis (SNA). These methods emphasize only on communication and the graph structure of social networks but do not consider interactions, user interests, and the effect of user influence on Internet social networks. Some recent attempts have shown that using node or edge content of social networks may be effective in discovering communities or important people in a network [1,21,26,27,35,41,55]. Some of the methods do not allow users to register in different communities, which is considered a problem. A few researchers also believe that in some applications, each node belongs to one community, but most applications require nodes to overlap. With the purpose of solving this challenge, researchers have suggested other methods based on the Bayesian probability model [8,14,23]; this model allows overlap of community members, but this method frequently focuses on the structure of network graphs.

Some of these attempts through exploiting operations of users discover influenced people on Internet social networks [1,21,35].

A number of works address community discovery with a very strong assumption: To be called a community, a group of vertices must follow a very strict structural property. This task is similar to the very well-known data mining problem in network analysis, i.e., graph mining. Some examples of graph mining algorithms are given in [4,31,38,50]. However, traditional graph mining algorithms only return all the single different structure patterns with their support. In community discovery, there is only one important structure and the desired result is the list of all vertex groups that constitute that structure in the network. The methods reviewed here are clique percolation [39] and its evolution for bipartite graphs [33], *s*-plex detection [29], and the maximal clique approach [45]. Other minor evolutions will not be highlighted, such as *k*-dense approaches [43].

A past approach to this definition can be found in the block model family of solutions. In particular, some works focus on the definition of ‘structural equivalence’ [5,15], where the authors have defined the notion of structural equivalence generally by looking at the pattern in the nodes’ connections: If they are connected to the same (or equivalent) network portions, then the nodes are in the same community because their 'role' inside the network is the same. Since a defined structure may be, with no constraint, overlapping, weighted, directed, or multidimensional, there is virtually no structural feature that cannot be rooted in a definition used by the algorithms in this class. Depending on the desired structure, analysts can also find communities that do not overlap with any of the previous categories, thus avoiding densities, or bridges, or any other previous definitions. The shortcoming of this strategy is in working in incremental settings; given a simple structure modification, such as adding or deleting a single node or edge, the algorithm is likely to re-compute everything from scratch. This is because substructure properties that are discovered may be disrupted by any single modification.

Palla et al. [39] suggested that a community can be interpreted as a union of smaller complete (fully connected) subgraphs that share nodes. The authors defined a k-clique community as the combination of all k-cliques that can be reached from each other through series of adjacent k-cliques. Two k-cliques are called adjacent if they share $k - 1$ nodes. A two-clique is simply an edge, and a two-clique community is the combination of those edges that can be reached from each other through a series of shared nodes.

An s-plex [29] is a reduced concept of the c-isolated clique [24,25]. Alternatively, the authors employed a relaxed version of a c-isolated clique called s-plex [3].

Bi-clique [33] is a bipartite graph version that solves various problems regarding the k-clique approach [39], namely the impossibility to analyze sparse network regions due to the fact that two-clique communities are merely connected network components. The algorithm starts by isolating the N maximal bi-cliques in the bipartite network [47]. Using this list, two symmetric clique overlap matrixes were created for the two node classes. Then, both the matrix's diagonal elements greater than or equal to a and b (the two algorithm parameters), respectively, are set to one, while everything else is set to zero. The final overlapping matrix is obtained by matrix intersection using the AND operator. The final step entails determining the connected components of L; each component corresponds to a bi-clique community. The ultimate approach complexity is $O(m^2)$.

EAGLE [45] starts from the following assumption: Every dense-linked community has at least one large clique. This clique may be considered the core of the community. EAGLE initially identifies all maximal cliques in the network with the Bron–Kerbosch algorithm [6] [complexity $O(3n^3)$] and discards those whose vertices are part of other larger maximal cliques and those with less than k vertices. EAGLE then calculates the similarity between each community pair. Subsequently, it selects the pair with the maximum similarity, incorporates it into a new community, and calculates the similarity between the new community and other communities. The similarity measure is known as the modularity [9]. This calculation recurs until only one community remains, thus completing a dendrogram (tree diagram). The second stage involves cutting the dendrogram. Any cut through the dendrogram produces a network cover. To determine the cut location, a measurement is required to ascertain the quality of the cover by computing with a given variant of modularity.

Troussas et al. [46] described the mining for relationships among user clusters on Facebook for tutoring languages. The K-means clustering algorithm was applied to determine groups of users with the same learning styles and capabilities. Information was extracted from the user dataset and transformed into a comprehensible structure for further use. The authors used seven user characteristics (such as age, sex) for clustering and analysis. It was shown that the Facebook characteristics selected seemed to be significant cluster determinants.

In another set of studies, the semantic content of a social graph was used to explore the communities involved. The community user topic (CUT) model is an instance of such work [52].

CUT similar models assume that people who actively talk about a specific issue (identical in relation to a particular issue) are connected. SSN similar models presuppose that users who are connected have similar interests. These two assumptions are not always true in real-life scenarios [42]. To address this challenge, a number of researchers have employed a combination of topology graphs and content (posts) to discover communities [40,51]. By using the content of vertices and edges in discovering groups, researchers have indicated that the content of vertices and edges can be useful in improving the quality of a community structure [41,53]. Different research has also emphasized ontology in exploring communities, and the concept of semantic communities and semantic link networks (SLN) has been introduced [55].

An approach is provided in [11] that recommends similar users, resources, and social networks to users by considering not only local information but also global information and four user action types (membership, friendship, posting, and evaluation). It operates on a social internetworking context and is based on the hyper-graph model.

A remote evaluation tool based on the usage data of users (logging information) is provided in [12] that identifies usage patterns based on client-side event logs. It employs the usage graph of users' actions and considers sequence alignment method (SAM) for measuring the distance between event streams. The work opens up new possibilities for the application of Web mining techniques and recommending similar users based on their actions on Web pages or on social networks.

In the current study, a method is proposed with the purpose of discovering communities. In addition to graph structures, social network content is used and the convergence of other communities is allowed. In this method, the important nodes are first discovered, based on which communities are formed. The following section refers to related attempts with the proposed method. Goyal et al. [21] presented an algorithm for selecting leaders through node influence on nodes with the assumption that social networks include one graph and one operation table, where the graph shows friendships among users and the operation table represents each user's actions. With the frequent pattern algorithm (SWF) in this method, Goyal et al. analyzed the operation table and wrote in a matrix. Ultimately, important people in the network (leaders) were identified. In this method, leaders discovered are people who have carried out a specific activity faster than their friends. A propagation graph was first proposed in their research, and through creating this graph for each activity or operation in the network, a user is identified as a leader. Two thresholds, confidence and genuineness, are considered in differentiating real from unreal leaders.

Adnan et al. utilized closed frequent patterns to discover communities in social networks. They assumed a set E including n nodes, $E = \{e_1, e_2, e_3, \ldots, e_n\}$, where each node e_j is linked to the dataset D_j that represents the operation relating to this entity or node. Thus, for each entity, there is one transaction data set. Through using closed frequent patterns, this set of data can be modified into a vector, which can be considered an effective representative of operation carried out through a related node. The vector similarity that indicates the operation of nodes is calculated through a similarity measurement such as dot product. Upon measuring the similarity and distance, it is possible to discover communities through standard clustering methods [41].

In some research, leaders in the network are first discovered, beyond whom a community is developed. For example, with the purpose of discovering community, Kanawati [26] first obtained leaders through the measures of centrality, degree centrality, betweenness centrality and closeness centrality. Then, neighbors related to each leader establish a small community such that typical nodes are ascribed to the nearest leader and the community relating to them is discovered. Khorasgani et al. [27] also found valid leaders and people on the basis of two centrality measures: degree centrality and betweenness centrality. Each non-leader node is reviewed in terms of how much similarity exists, so that each node's ratio to each leader has one weight. After voting among neighbor nodes, the nodes are attributed to one leader. In this way, there is a community with one leader, where the leader is found through the above-mentioned metrics and each node is reviewed in order to determine the leader. These steps continue until convergence is reached. The *K*-means clustering algorithm has motivated this approach. For estimating the *K* value, some methods are suggested in [27]. Lu et al. [35] researched the Flickr social network and similar sharing service networks, and by using the factors of favor volume, favor converge, and favor timeliness, valid people (leaders) are discovered. Favor volume means pictures liked by other members. Favor converge shows the spread of one member; this not only means the fans of one member but is considered the degree of the influence of fans. Favor volume is viewed as complementary in preventing fans' prejudice toward one member's pictures. People who have influential fans have greater weight for this factor. Favor timeliness refers to the time sensitivity of fans. A member with high favor volume, favor converge, and the most recent fans probably has high validity and skill in the network. The evolution challenges in social networks following which valid people will change were considered in Lu et al.'s research, and a solution was proposed.

These methods merely concern discovering valid people, while social network content is not important in discovering communities. As an example, in Kanawiti's [26] method, a leader is discovered through degree and betweenness centrality, and then, neighbors related to each leader establish a small community.

3 Proposed method

3.1 Assumptions

A social network can be a simple graph consisting of nodes and edges. For example, consider a social graph $G = (V, E)$ that includes six users $V = \{U_1, U_2, U_3, U_4, U_5, U_6\}$. The relation between users is shown with edges in Fig. 2. In this study, it is assumed these relationships signify friendships between users.

Also, any of the users in the network can perform some or all of a set of permitted operations. The allowed operations on social networking sites can differ. For instance, on the Web site Good Readers, users can add their own books to the shelves, check, and rate books. They can view what books their friends are reading and discuss and debate specific issues. Users can receive various suggestions from friends. Selecting books by users can be considered a user operation. Table 1 shows an operation table of operations performed by users.

Fig. 2 A simple social graph

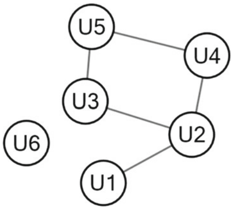


Table 1 Sample operation table for users on a social network

User	Actions
User1	Action1, action3, action5, action6
User2	Action2, action3, action7
User3	Action3, action5, action6, action7
User4	Action2
User5	Action3, action7
User6	Action1, action5

3.2 Steps in the proposed approach

A new approach is proposed for the discovery of communities by performing frequent pattern mining [41] on the operation set of users and using a classification algorithm on communication among users. The method is tested on social network Web sites. The assumption is that a community consists of a number of leaders and followers, as described in detail later. In this method, an attempt is made to obtain small groups of users in such a way that users from each group are similar in terms of performance on the network. Then, users who are neighbors in this sense are discovered as followers of these groups. Subsequently, a summary of the steps in this method is presented. This method to discover communities in social networks consists of four principle stages (Fig. 3):

1. Data preprocessing;
2. User frequent pattern mining (obtaining harmonious groups);
3. Confirmation of harmonious groups as small communities;
4. Expanding a small community.

- Step (1) If necessary, preprocessing will be done on the dataset to provide the inputs including the user operation table and neighbor table according to the algorithm. In this step, three algorithm thresholds regarding network information is estimated. The thresholds α, β and ¥ will be explained subsequently.
- Step (2) The frequent pattern mining algorithm is run on the operation user table to achieve the maximal patterns. Each pattern includes a sequence of users having a similar pattern in the network. Each user sequence is named a homogeneous group. ¥ is used as a minimum support of frequent pattern mining in this step.
- Step (3) Users in each homogeneous group reveal people with similar performance and taste in the network, but people in a group spread over the social graph and a dense group can be established; people in these groups can act as members of the group if they are linked to each other. This link does not only mean a direct connection or the existence of edges among nodes, but also mean communication with mediator nodes accepted in a threshold, in which case this is the β threshold. The outputs from this step is a group of users, in which similar and related users are its members. Such groups are called small communities.
- Step (4) A small community as identified in the previous step covers a small number of users. This step aims to expand such obtained communities. Each small community is viewed as a core of one community, and neighboring people are considered followers of these small communities. Thus, small communities are expanded by taking into

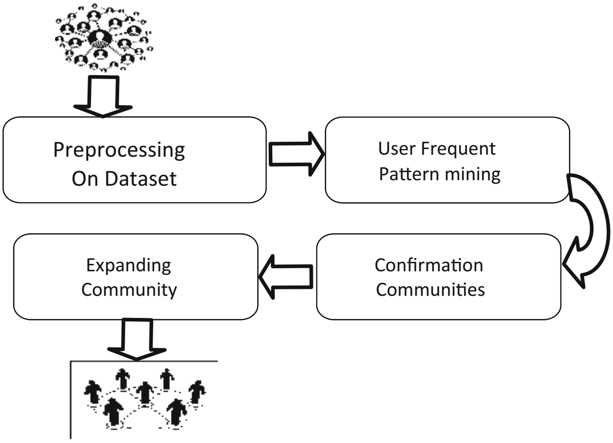
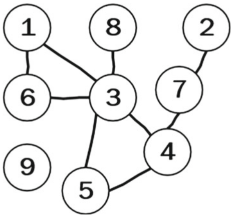


Fig. 3 Community discovery steps in the proposed method

Fig. 4 Friendship relation among users



consideration communication and changes into acceptable community sizes. The α parameter is used as a threshold in this step.

From the explanation of the proposed method steps, it is concluded that ultimately discovered communities include users, who besides sharing behavioral similarity are close to each other on a social graph. These steps are explained below in more detail.

3.2.1 First step: data preprocessing

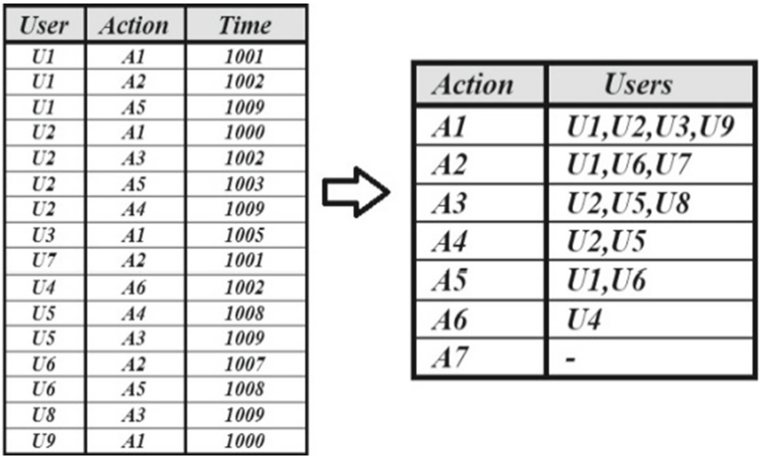
As input of the proposed method, an operation user table and neighbor table (graph) are required. The operation user table shows user operations regarding actions in a network, and the neighborhood table includes friendship communication among users (Fig. 4). This table is normally displayed as a graph.

- Graph of friendship among users

- Operation user table

The currently proposed framework focuses on social network Web sites. On these Web sites, there are normally sets of social operations, whereby each user can influence the operations with their individual decisions. On networks like Facebook, Twitter, and YouTube, for instance, various views about allowed operations are considered. It is worth noting that it is important for allowed operations in the network for users to be free of charge. Operations such as login and logout, for example, may not be deemed operations in the operation table, because all users perform these.

Table 2 Creating an operation user table from the operation file



User	Action	Time
U1	A1	1001
U1	A2	1002
U1	A5	1009
U2	A1	1000
U2	A3	1002
U2	A5	1003
U2	A4	1009
U3	A1	1005
U7	A2	1001
U4	A6	1002
U5	A4	1008
U5	A3	1009
U6	A2	1007
U6	A5	1008
U8	A3	1009
U9	A1	1000

Action	Users
A1	U1,U2,U3,U9
A2	U1,U6,U7
A3	U2,U5,U8
A4	U2,U5
A5	U1,U6
A6	U4
A7	-

Table 2 shows an operation file that is usually in triple form (user, action, time). For the purpose of building an operation user table, we use columns of users and actions.

- Goals of the first step
  - Preparing a neighborhood table (usually does not need preprocessing).
  - Preparing the user operation table.
  - In addition to providing the inputs to the proposed method, the threshold values of α, β, and ¥ are estimated in this step. It is preferable to estimate these thresholds through expert and in terms of the social network size, number of nodes, number of edges, and number of operations in the network.

3.2.2 Second step: user frequent pattern mining

Generally, a frequent pattern refers to a pattern of data that seems to repeat frequently in a set of special data. Exploring such patterns provides a lot of usage and different kinds of algorithms, which have specific advantages and disadvantages. After preparing the input in the first step, frequent pattern mining is run on the user operation table. With the purpose of minimizing time complexity in this step, it is ideal to use a parallel algorithm of frequent pattern mining or some algorithm based on the prefix tree.

- Performing the step of user frequent pattern mining

Assume that the algorithm of frequent pattern mining is performed on the table of user operation (Table 3). The output is a sequence of users that was observed in a similar operation (equal to the threshold or above). Regarding minimum support of ¥, the algorithm for discovering frequent patterns is performed on this table and the maximal pattern is explored [16]. It is worth mentioning that the maximal patterns with single item will be pruned. The maximal pattern in Table 3, if $¥ = 2$, will be $\{U_2, U_5\}$, $\{U_1, U_6\}$, showing that these two users have similarities in performing 2 or more actions. Frequent pattern mining algorithms such as Apripri [2] can be used.

Table 3 Performing the algorithm of frequent pattern mining on the operation user table

Frequent pattern	Maximal pattern
U1	U1, U6
U2	U2, U5
U5	
U6	
U1, U6	
U2, U5	

Minimum threshold ¥ Each algorithm for exploring frequent patterns needs minimum support. According to this parameter, it is a frequent pattern whose number of repetitions in the dataset is larger than or equal to the minimum support. Here, this threshold plays a significant role that users will be explored with how much number of similar action as a pattern. It is recommended for social network experts to estimate this threshold regarding the number of operations in the network, because it directly affects discovered communities.

- Goals of the second step.
  - The output of this step after performing the algorithm of frequent pattern mining is a sequence of users, which is hereby named a homogeneous group. Each homogeneous group consists of two or more users, showing that users in this group perform similar operations according to threshold ¥ or above. Thus, it can be said that users in each group are similar in terms of the operations they perform.

3.2.3 Third step: confirming small communities

It is obvious that a homogeneous group includes users who are comparable to each other regarding operations performed, but these users may have a high degree of separation on the social graph. In other words, users in a group are spread apart. It is better for the members in a group to be related to each other on the social graph in order to verify it as a dense group. In this step, the aim is to omit a non-dense group into more group and confirm dense groups. By a dense group, it means that users are related to each other. Nonetheless, this relation does not mean a direct connection among nodes (existing edge) in all networks, but this represents the concept of weak and strong ties in a social network. A threshold of β determines the allowed numbers of intermediations.

Threshold β This threshold defines there are more edges between two nodes in the social graph for connecting two nodes. An expert determines this β threshold value. This value is not constant because there are networks of different sizes. By running the proposed method on an actual dataset and testing it, the most appropriate value of β is lower than or equal to 3 ($\beta \leq 3$).

In this stage, if users in each sequence obtained from the previous step reach a threshold of β, it will be verified as a small community; otherwise, this group will be reduced or split into smaller groups. The output of this step signifies groups of users who, at a ratio of ¥ or greater, have similar characteristics. On the social network graph, these users are at distances according to the number of β, and such groups are called small communities. The question considered here is how this theory will be established with minimum time complexity because there may be additional paths between two nodes on a graph. To attain

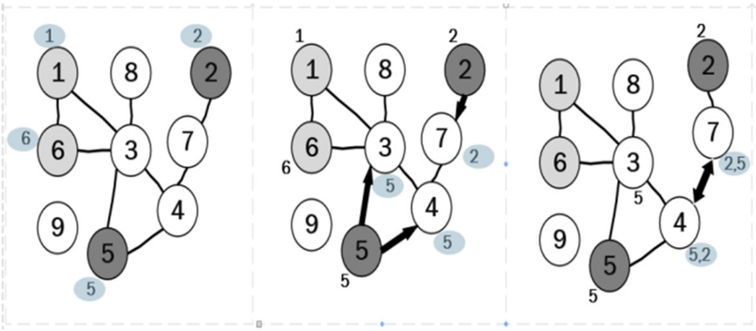


Fig. 5 Verification steps for a homogeneous group as a dense community

the purpose of this step, all paths between two nodes must be searched in order to find the shortest path. If the distance of the shortest path between two nodes is less than or equal to β, the group is substantiated. The pattern produced in the second step can be consistent with the groups in the networks. Also, there is no limitation to the length of the sequences obtained with the algorithm of discovering frequent patterns. The aim is therefore to find the shortest path among nodes on a weightless graph that has the lowest time complexity.

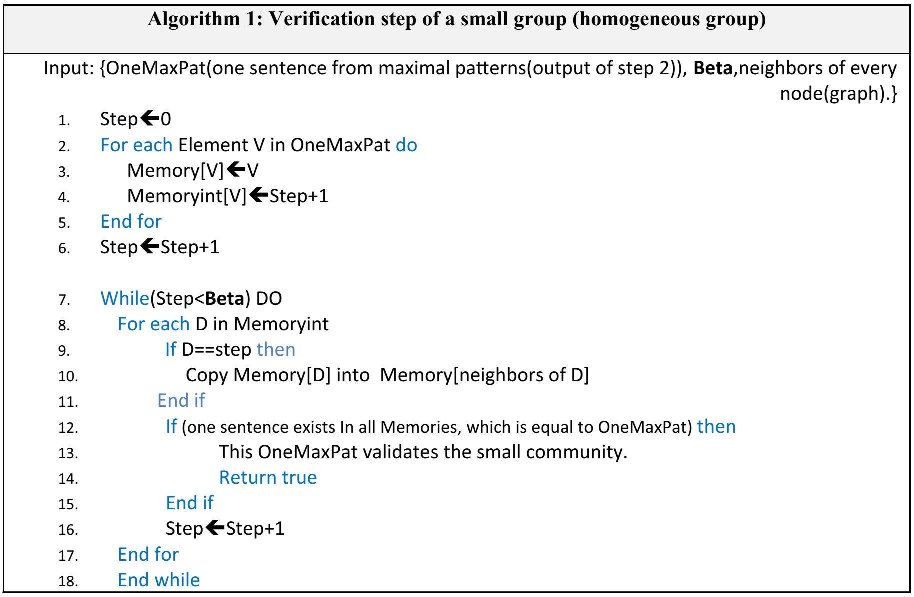
- Performing the verification step of a small group (homogeneous group)

The proposed solution offers simplicity and low time complexity and is additionally able to analyze the correctness or wrongness of connections among nodes in one run. This method does not store intermediate nodes, and this step necessitates the correctness or wrongness of nodes' linkages regarding the β threshold. To better understand this stage, consider the steps presented in Fig. 5.

With this example, the goal is to review the correctness of the relation between nodes 2 and 5 (which were extracted as a homogeneous group in the previous step) with respect to the β threshold. Assuming $\beta = 2$ in this example and if these two nodes were confirmed as one group in this step due to the shortest path, then there are two intermediate nodes. In the first step, each node belonging to the group will save its name in its memory, while the memory of other nodes in the first step is empty. In the second step, all nodes belonging to the group will convey their memory to their neighbors. The neighbors will receive messages and integrate them with their own memory. In this step, there is the variable 'step,' and according to each sending stage, one unit will be added to it (see Algorithm 1). The maximum allowed value for step is equal to the β threshold value. After 'step' reaches this value, the nodes' memory will be reviewed, and if the memory of one node includes nodes in the group (here nodes 2 and 5 are group members), these nodes relate to the β threshold. Thus, in step 2, the 'step' value is equal to 1. Then, through reviewing the nodes' memory in the third step, the correctness of the two nodes' relation with thresholds with value 2 is validated. If there is no complete sequence in the memory of each node, the group will be omitted. For example, if $\beta = 1$, the group consisting of nodes 2 and 5 will be cut. These steps are also run for another small community discovered (here, nodes 1 and 6).

With the purpose of optimizing the above-mentioned algorithm, it is better that after sending, the memories are reviewed if a complete sequence is found. This means that the relation between nodes linked to each other and intermediated nodes is lower than the β threshold. Therefore, to prevent processing overflow, the procedure is stopped. However, with each sending, it is not logical to review the memory of all nodes, even nodes with an empty memory. So through bitmap and equaling every bit relating to the nodes whose memory is changed to 1, the nodes engaged can be reviewed and high processing is avoided. Generally, the algorithm has 2 conditions of whether there are complete sequences in the nodes' memory or the 'step' value reaches the threshold β value.

The pseudo-code of this method is displayed below.



Algorithm 1: Verification step of a small group (homogeneous group)

Input: {OneMaxPat(one sentence from maximal patterns(output of step 2)), **Beta**,neighbors of every node(graph).}

```
Step←0
For each Element V in OneMaxPat do
    Memory[V]←V
    Memoryint[V]←Step+1
End for
Step←Step+1

While(Step<Beta) DO
  For each D in Memoryint
        If D==step then
            Copy Memory[D] into  Memory[neighbors of D]
        End if
        If (one sentence exists In all Memories, which is equal to OneMaxPat) then
                This OneMaxPat validates the small community.
                Return true
        End if
        Step←Step+1
  End for
  End while
```

- Goals of third step:
  - To have a correct relation among members regarding the β threshold, each group consisting of two or more nodes is reviewed. If there is a correct relation among members, the group will undergo the next step without change; otherwise, it will be omitted.
  - The output of this step represents groups of users, where the users in a group share similar network features equal to or higher than the number of ¥ and are on a social network graph with distances according to the number of β. These groups are called small communities.

3.2.4 Forth step: expanding small communities

Each small community can be considered a set of leaders (people who are more skilled in certain areas, or in a social network they affect people and are thus known as leaders [35]). Leader has a role in some network operations, and he plays as an extender of it. Users who are in the neighborhood of a small community are likely similar due to the close relation with people in the community and are directly influenced by their neighbors. For instance, assume that a user has a role in an Internet social network. This user's friends are able to observe that user's actions, which is an opportunity for the friends to carry out those operations themselves. Now assume that among these people some perform such operation, common friends of these people communicate with each other and are thus eager to perform those actions [21]. This spread of actions is also the basis for virus marketing on the Internet. Thus, with all allowed operations in a network and people having similar taste network operations, masses of people are formed who spread such similar operations. It can therefore be said that by expanding small communities, similar people or people acting toward this will be discovered.

It is assumed that the neighbors of a small community will follow it, and so the community discovered in a previous step will be developed. Also, according to the α threshold, communities of acceptable sizes will be obtained. As such, an algorithm is used, which is similar to categorization algorithms. Nodes with no specific community are attributed to the closest small community as followers of that community. In other words, each node of a non-leader is assigned to the community that has more votes from that community. In the following sections, we refer to the proposed method to spread communities.

- The step of expanding small communities

After the third step, small groups of users are extracted as small communities. Each small community includes nodes that are alike regarding operations performed in the network. As mentioned before, people may be influenced by their neighbors, and if more neighbors of one node perform similar operations, the influence is greater. This method uses two thresholds, α and μ in order to spread the identified groups.

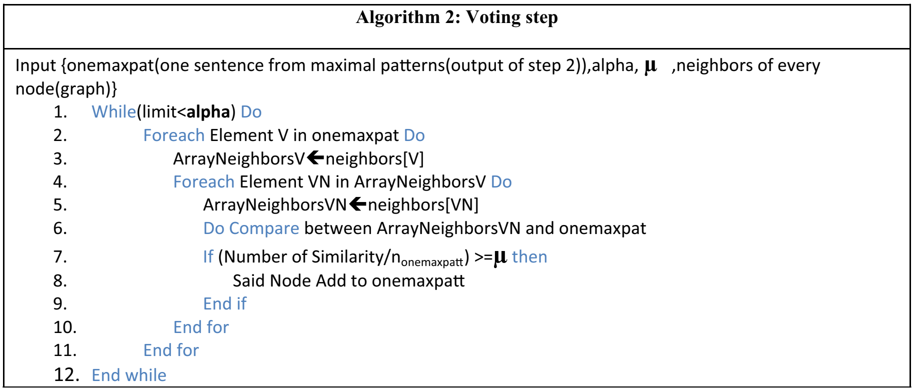
μ threshold This threshold in the algorithm determines how many neighbors of one node belong to the community, so as to discover the above-mentioned node as belonging to the community. The value of this threshold is between 0 and 1.

α threshold This threshold shows which level of people belonging to a community can accept neighbors as their followers. The larger the value of this threshold, the higher the number of people belonging to the community after development.

- **Implementation**

In this method, usual nodes (nodes that are not leaders) are attributed to a community that has many neighbors of nodes belonging to a specific community. The community may overlap, and also, some nodes that are not attributed to that specific community after voting are deemed outliers.

The pseudo-code of the algorithm (Algorithm 2) with an example in Fig. 6 to understand this method is provided below:



Algorithm 2: Voting step

```
Input {onemaxpat(one sentence from maximal patterns(output of step 2)),alpha, $\mu$ ,neighbors of every node(graph)}
While(limit<alpha) Do
    Foreach Element V in onemaxpat Do
        ArrayNeighborsV🡐neighbors[V]
        Foreach Element VN in ArrayNeighborsV Do
            ArrayNeighborsVN🡐neighbors[VN]
            Do Compare between ArrayNeighborsVN and onemaxpat
            If (Number of Similarity/$n_{onemaxpatt}$) >=$\mu$ then
                Said Node Add to onemaxpatt
            End if
        End for
    End for
End while
```

$n_{\text{onemaxpatt}}$ shows the number of people in the onemaxpatt community.

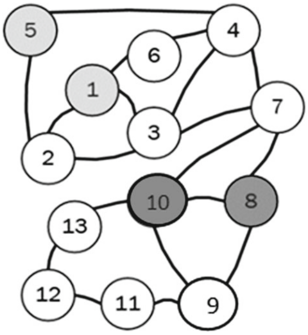


Fig. 6 Voting method

To better understand this algorithm, consider 2 small communities, $G_1 = \{1, 5\}$, $G_2 = \{8, 10\}$ in the social graph of Fig. 6. Both communities consist of two members on the graph.

If the threshold is $\mu = 0.6$, then for a user to be a member of a community of G_1 and G_2, they need to be neighbored with 60 % of nodes in these communities that means neighboring with each of two nodes in step one, because these two communities in the first step consist of only two nodes (Fig. 7).

- In the first step where $\alpha = 1$, nodes 7 and 9 link to group G_2 and node 2 links to group G_1.
- In the second step where $\alpha = 2$, node 3 is linked to group G_1 [here 60 % of neighboring for G_1 that consisting of three nodes (node 1, node 5, node 2), is two nodes] because node 3 is linked to all two nodes in G_1, group G_2 will be unchanged because non-nodes are linked to 60 % of group G_2 (node 8, node 9, node 10, and node 7)
- In the third step where $\alpha = 3$, node 4 is not linked to group G_1 because this node is not related to 60 % of the nodes in the group, and this group remains unchanged.

With the purpose of discovering a community, the proposed method is aimed at discovering similar nodes as a core (leader) of a community. This core will be expanded on the social graph through communication and data expansion to cover more nodes as a community (Fig. 8).

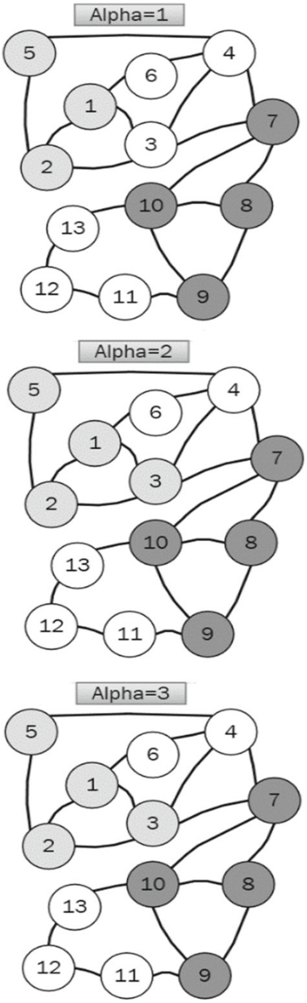
- Goals of the fourth step.
- In this step, the fundamental goal of the proposed method is attained, meaning that communities are obtained. The communities include nodes whose operations are alike.

4 Evaluation

Different methods for discovering community structures create problems in the evaluation of these methods. Evaluating an algorithm means confirming the considered algorithm in order to solve a specific problem in a method. With this purpose of discovering communities in the structure of social networks, it is obvious that all algorithms aim to identify similar nodes as a community. Among the fundamental challenges in evaluating methods of obtaining communities is that there is no explicit definition of community structures in real-world networks and each algorithm calculates the similarity among nodes regarding their usage.

One of the main methods of evaluation is a benchmark standard graph. Such graphs are created to evaluate community discovery. The existing nodes in these graphs have class labels, meaning the nodes are grouped ideally and researchers can evaluate their work with

Fig. 7 Voting and community expansion



these graphs. Some graphs are produced through standard computer methods, and others are compiled from real-life social network data.

4.1 Evaluation metrics

Social network experts have introduced a number of standard evaluation metrics that can serve evaluation purposes. One of these metrics is the *F*-measure, and it is employed for evaluation in the current study. This metric includes two fundamental factors, i.e., precision and recall, which are obtained from the following relations [49].

(1) $P = \dfrac{\text{cells correctly put into a cluster}}{\text{total cells in cluster}}$

(2) $R = \dfrac{\text{cells correctly put into a cluster}}{\text{All the cells that should have been in the cluster}}$

(3) $F\text{-measure} = \frac{2 \times P \times R}{P + R}$

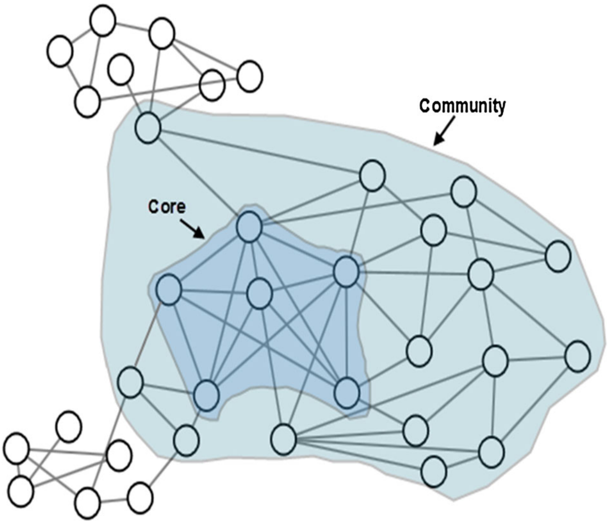


Fig. 8 Creating a community in the proposed method

4.2 Dataset

Two social network datasets are utilized for assessment purposes: Flickr and BlogCatalog.

4.2.1 Flickr dataset

Flickr is one of the largest sites for video and picture sharing, Web services, and Internet communities, created by Ludicorp in 2004 and bought by Yahoo in 2005. A user in this social network can share pictures and be a member of different groups. Any member on this site can create a Flickr group. The creator of a Flickr group can control and determine the group's limitations. Groups are communication links among Flickr members beyond videos and pictures. Sometimes, a user receives private message when is part of a group. Users can label videos and files that have been shared. A large dataset is spread from this network. The dataset employed for this article was gathered in 2012 and includes more than 35,000 users. Due to having user grouping files (class labels), this dataset is appropriate to evaluate the algorithms of discovering communities and classifications. It includes the following data:

- Relations: information of (undirected) friendship relations among users.
- Content created by users: labels of users' pictures are gathered.
- Groups in which a user is a member are stored.

4.2.2 BlogCatalog dataset

BlogCatalog is a social network and one of the best tools for finding Weblogs and Weblog designers. Users are able to introduce their Weblogs and save public and private notes. Moreover, users can have friends in this network. The BlogCatalog dataset applied in this research includes 90,000 users. Because this dataset includes user grouping information, it is appropriate for the evaluation of algorithms for community discovery and classifications. This dataset includes the following information:

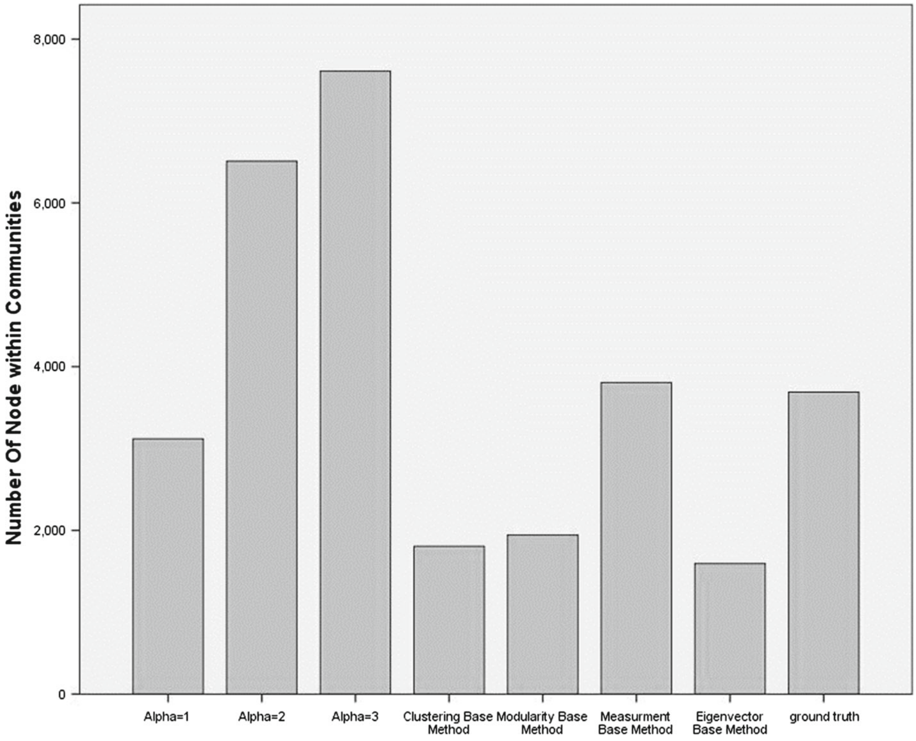


Fig. 9 Average number of people in the community using different methods of community discovery on the Flickr dataset

- Relations: information of (undirected) friendship relations among users.
- Content creation by users: labels considered by users for their Weblogs.
- Groups: each user is grouped according to their Weblog labels.

4.3 Results

By performing different methods of community discovery on two datasets, namely Flickr and BlogCatalog, communities were discovered and the results are presented in Figs. 10 and 12 (for the Flickr and BlogCatalog datasets, respectively).

The two-dimensional bar graph in Fig. 9 shows the average number of people in a community with different methods for the Flickr dataset. The 8th bar is the average number of people in an ideal grouping. As mentioned in the introduction for this dataset, the dataset includes user grouping information. In the grouping, 201 separate groups were discovered with an average of 3688 group members.

As seen in Fig. 9, the higher the value of α in the proposed method, the higher the average number of people will be.

With the implementation of various methods on the Flickr dataset, communities are detected and the results are shown in Fig. 10. The average precision, recall, and F-measure metrics values are shown as a diagram. Each bar shows a separate community detection method. The first three bars are related to the proposed approach with different α values. The fourth bar represents modularity-based community detection [9], the fifth bar is for the

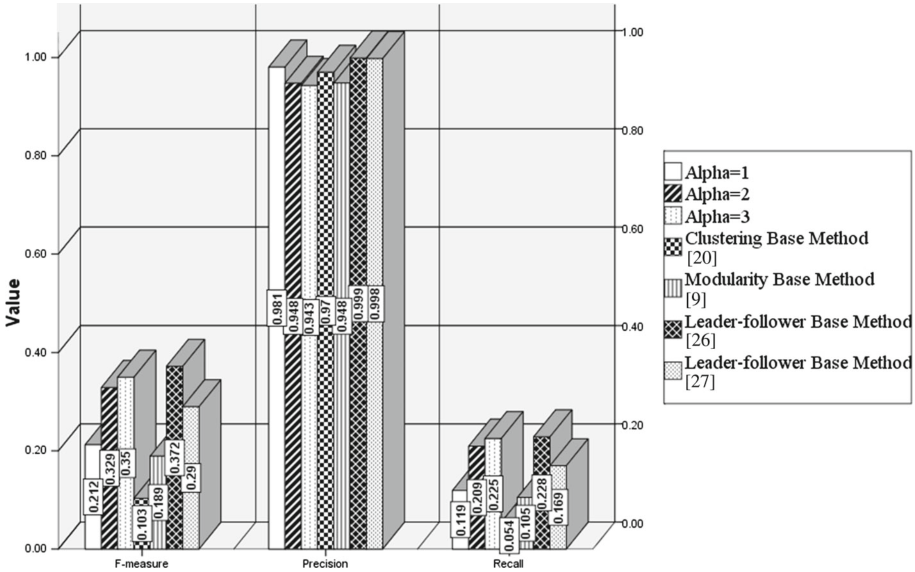


Fig. 10 Average of *F*-measure, precision, and recall in different methods of community discovery on the Flickr dataset

clustering-based method [20], and the sixth and seventh bars are for the leader–follower-based method [26,27].

The reason for selecting the two methods based on clustering and modularity for comparison with the proposed method is to show to what extent using content in the proposed method is effective in detecting communities. In these two methods, only the graph structure is used to detect communities, but both leader and follower-based methods (sixth and seventh bars) act like the proposed method, where the leader nodes are first identified based on network criteria and then communities are extracted around each leader. The difference is that in the proposed method, a set of similar users is detected as a community leader based on their actions, while in the other two methods [26,27], the network leaders are detected from the network's graph structure. Each community has only one leader, and around these leaders, communities are detected.

A summary of results from the Flickr dataset (Fig. 10) signifies that the proposed method is more flexible than other methods, because with different values of threshold α, different values are obtained for precision, recall, and *F*-measure. In implementing the proposed method with a value of $\alpha = 1$, community recognition accuracy is more than with other α values. In this case, the number of people in a community is less than $\alpha = 2$ and $\alpha = 3$. Figures 9 and 10 show that the precision and recall values are directly related to the average number of people in the community. The higher the number of people in the community, the higher the value of α will be, and if the precision value does not decrease significantly, the *F*-measure will increase. Clearly, the proposed method with this dataset for the *F*-measure is superior to other methods (except one approach) (Fig. 10).

Next, the results from the BlogCatalog dataset are compared. As mentioned earlier, this dataset includes user grouping information. Here, 319 separate groups were discovered, and each group had an average of 842 users as members (groups with at least ten users) (Fig. 11).

According to the BlogCatalog dataset results (Fig. 12), although the flexibility of the proposed method is confirmed, it is observed that this method's accuracy of community

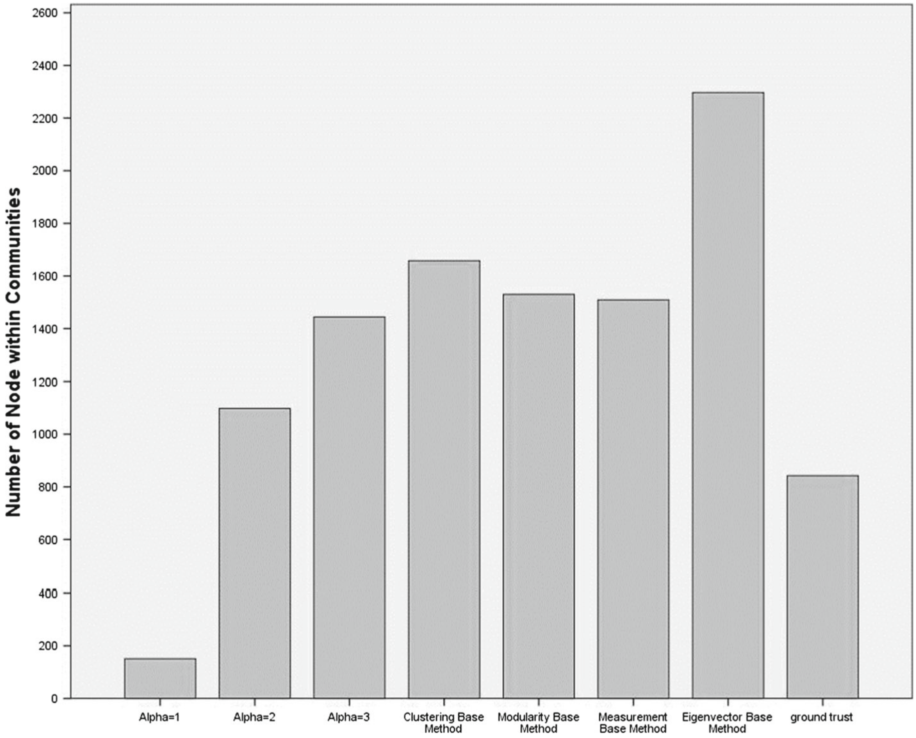


Fig. 11 Average number of people in the community using different methods of community discovery with the BlogCatalog dataset

discovery on some social networks is higher. On the other hand, the proposed method run with this dataset is superior in terms of the *F*-measure value than other methods (except one).

It can generally be said that the proposed method produces different dataset results regarding the input parameters. It was shown that with increasing flexibility (adjusting parameters according to needs) the quality of the community enhances with the proposed method (*F*-measure includes accuracy and recall).

In attempting to provide a way to solve the problem, researchers need to compare their proposed methods with other studies in the same field. Despite the many studies and extensive research on discovering methods of evaluation, the evaluation of these algorithms remains an open question, which should be based on network structure analysis. On the other hand, well-known techniques of exploring the structure of a community often only consider the social network graph and our approach is different from these methods. Thus, comparisons with their results are not suitable and are withdrawn.

5 Time complexity of the proposed method

In this section, the time complexity of the proposed method and other methods' is considered. The symbols related to proposed method are listed in Table 4.

In Table 5, the results of calculating time complexity and other parameters are shown. According to the results, it can be deduced that the proposed method has higher time com-

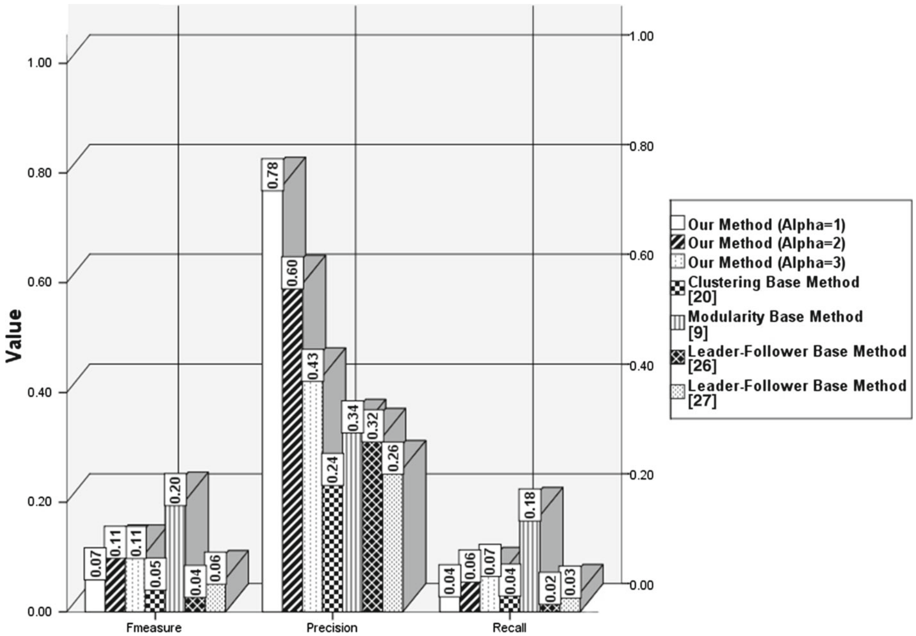


Fig. 12 Average of *F*-measure, precision, and recall using different methods of community discovery with the BlogCatalog dataset

Table 4 Symbols

Symbol	Description
n	Number of vertices of the network
m	Number of edges of the network
k	Number of communities of the network
T	Number of action in the network
A	Max number of actions for a node

plexity than other methods, because in addition to graph structure, this method employs user content to increase community discovery accuracy.

The use of iterative algorithms renders the proposed method more complex than other methods. The content structure-based approach that combines information from the graph and content to discover communities aims to increase the accuracy of communities that have been found. However, because complexity is undeniably one of the most important parameters in detection methods for assessing community detection, pruning nodes and using different iterative algorithms to explore recurring patterns can be used in future works to optimize the time complexity and running time.

6 Advantages and disadvantages of the proposed approach

All community detection methods have advantages and disadvantages. The following are some of the advantages and disadvantages of the proposed method:

Table 5 Comparison of time complexity

Method	Use content and graph structure	Overlap	Outlier	Complexity time	References
Clustering base method	–	–	–	$O(n^3)$	[20]
Modularity base method	–	–	–	$O(mk\log_n)$	[9]
Frequent pattern algorithm base method	✓	✓	–	$O(TAn^2)$	[21]
Leader–follower base method	–	✓	✓	$O(n^2m)$	[26]
Leader–follower base method	–	✓	✓	$O(n^2m)$	[27]
Our method	✓	✓	✓	$O(n(T^2+k+m))$	–

Advantages:

- *Improved quality of community* The proposed method extracts users who are similar in terms of operations done on the network. If the users have a close relationship on the graph, they are identified as a homogeneous group. Users who are neighbors of a group may be affected by the same operations of the group and even tend to do the same actions. So, a homogeneous group extents with its neighbors and creates almost convergent communities. Finally, a community is created that consists of a number of leaders and followers.

  Tests show that community mining using the proposed method is acceptable in real-life situations.
- *Layer communities* Several methods have been proposed to detect communities and discover network leaders. In the method proposed in this work, according to the scores for nodes within each group following grouping (community detection), nodes with different priorities and hierarchies are assigned to groups. It also becomes possible to extract strong communities.
- *Parameters* Parameters are set to detect suitable communities according to specific applications. Input parameters allow network professionals to determine the number of communities, the number of nodes within the community, and the number of overlapped communities and outliers according to the application employed.
- *Overlapping* Community detection methods often do not allow users to be members of different communities, which can be problematic. Although some researchers believe that it is better for some applications if each node belongs to only one community, the majority of applications require overlapped communities. To solve this challenge, researchers have proposed Bayesian probability modeling. These models allow communities to overlap. In the presently proposed method, a person may belong to different communities.
- *New approach in community detection* A new approach to social network analysis was introduced that is based on data mining tasks and the use of both user-generated content (user actions) and the relationships between users in community detection. We hope this approach will be useful for researchers in this field and motivate new ways in this field.

Disadvantages:

- *Time complexity* The complexity of the proposed method is totally dependent on the input parameters. Although the main objective of this research was to improve the quality of communities using the characteristics of users in specific networks, complexity is recognized as an important criterion in the evaluation of community detection algorithms. The complexity of the proposed method will be higher in most cases than in other methods, despite the iterative algorithms and user properties in the community discovery.
- *Parameters* Determining the appropriate parameters is often done through trial and error. Working with these methods is more difficult than with other nonparametric methods. However, this feature can also be deemed an advantage. But one of the characteristics of an algorithm is the lack of input parameters. An algorithm should be able to deduce explicit knowledge without the need for any additional information.

7 Conclusion

The increasing data availability on social networks has motivated computational research on social network analysis. Recently, community discovery in social networks has become one of the most significant challenges in social networks. In this paper, a new method was proposed on the basis of personal interests of users and their social relations in order to discover communities. There are two general methods of discovering communities on social networks, i.e., methods that discover communities on the basis of the relations among network users (methods based on the graph structure) and methods that are based on common interests of users in a network (methods based on content); second methods measure similarity of users' interests in social networks. Most methods of community discovery take into consideration one of these aspects. In fact, communication or content for obtaining communities in social networks is very important. The proposed method is a hybrid technique that considers content and graph structure in order to obtain a community. Through evaluating the proposed method on two real datasets, it was demonstrated that the proposed method is more appropriate than other methods of community discovery and is also more flexible. Basically, a new approach of analyzing social networks that discover communities on the basis of data mining and users' interest was proposed to improve community quality and to be consequently applied for friend suggestions, customer segmentation, and analyzing the effectiveness of specific networks.

All methods of community discovery have particular advantages and disadvantages. Advantages may be improved accuracy and quality of community detection, discovering leaders and communities, adjusting parameters in order to discover appropriate communities for specific usage, recognizing convergence and outlier nodes and new ways of discovering communities. The proposed method is also faced with challenges and disadvantages, such as time complexity due to using the frequent pattern mining algorithm and problems with determining parameters. We hope that this method is a new feasible way of discovering communities on the basis of content structure.

For future attempts, it is recommended to use the keeping algorithm of frequent pattern instead of frequent pattern mining such as CAN and CAT algorithms [16] that are potentially good alternatives in algorithms like Fp-growth and a priori for dynamic/online data, so this step will make the method flexible in discovering leaders using incremental data. It is also possible to increase the quality of results by changing the voting approach in the fourth step of the algorithm. To reduce the time complexity, using parallel ways should also be considered.

Appendix: Structure-based evaluation

We compared our detected communities in terms of structural metrics to other approaches [26, 27].

To demonstrate that the proposed method extracts the consistent community in term of structure and density, we have implemented the proposed method and also two approaches [26,27] on Last.Fm data set. For approaches in [26,27], we calculated core communities, each time with different centrality measures (are shown in method column) While in two papers [26,27], only betweenness, closeness, and degree metrics were mentioned, communities are formed around the cores in each method and are determined by voting from its neighbors (best results are shown from two approaches [26,27]). We compared our approach with them [26,27] in terms of structure metrics. Results are shown in Table 6 based on density, diameter and distance. As you can see, even the results of group analysis in our approach are not far from the previous works [26,27] in terms of average of density, average of distance between nodes, and average of maximum distance (diameter). Results indicate that our approach extracts communication along with similar users and somewhat related. Note the overlap is permitted in each method. This makes the average of density decrease, and average of distance and diameter increase. However, it is intended for all methods. It

Table 6 Structure-based evaluation

Method	Density	Distance	Diameter
Degree	**0.033**	**3.081**	**6.9**
Betweenness	0.023	3.184	7.2
Closeness	0.024	3.237	7.6
Eigenvector	**0.029**	3.293	7.8
PageRank	0.023	3.184	7.4
Our approach	**0.029**	3.426	7.6

should be said that the upper the density and lower distance and diameter is better but our approach is not far from the previous approaches, and this shows that our approach is acceptable in extracting the consistent community in term of structure and density; however, our aim is discovery of similar people in terms of performance and relationships in online social networks and is limited to a particular structure and its specific applications.

References

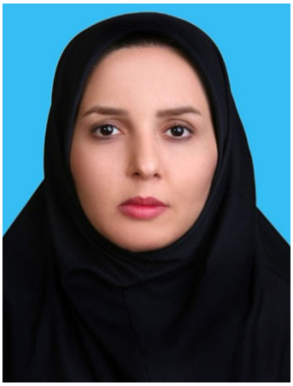
1. Adnan M, Alhajj R, Rokne J (2009) Identifying social communities by frequent pattern mining. In: Paper presented at the 13th international conference information visualisation
2. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: Paper presented at the proceedings of the 20th international conference on very large data bases
3. Balasundaram B, Butenko S, Hicks IV (2011) Clique relaxations in social network analysis: the maximum k-plex problem. Oper Res 59(1):133–142
4. Berlingerio M, Bonchi F, Bringmann B, Gionis A (2009) Mining graph evolution rules. In: Paper presented at the joint European conference on machine learning and knowledge discovery in databases
5. Borgatti SP, Everett MG (1992) Graph colorings and power in experimental exchange networks. Soc Netw 14(3):287–308
6. Bron C, Kerbosch J (1973) Algorithm 457: finding all cliques of an undirected graph. Commun ACM 16(9):575–577
7. Charu C, Aggarwal R (2011) Social network data analytic. Springer, Berlin
8. Clauset A, Moore C, Newman ME (2008) Hierarchical structure and the prediction of missing links in networks. Nature 453(7191):98–101
9. Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066111
10. Coscia M, Giannotti F, Pedreschi D (2011) A classification for community discovery methods in complex networks. Stat Anal Data Mining 4(5):512–546
11. De Meo P, Nocera A, Terracina G, Ursino D (2011) Recommendation of similar users, resources and social networks in a social internetworking scenario. Inf Sci 181(7):1285–1305
12. de Santana VF, Baranauskas MCC (2015) WELFIT: a remote evaluation tool for identifying Web usage patterns through client-side logging. Int J Hum Comput Stud 76:40–49
13. Dinh TN, Xuan Y, Thai MT (2009) Towards social-aware routing in dynamic communication networks. In: Paper presented at the IEEE 28th international performance computing and communications conference
14. Eliassi-Rad T, Henderson K, Papadimitriou S, Faloutsos C (2010) A hybrid community discovery framework for complex networks. In: Paper presented at the SIAM conference on data mining
15. Everett MG, Borgatti SP (1996) Exact colorations of graphs and digraphs. Soc Netw 18(4):319–331
16. Feng M, Li J, Dong G, Wong L (2009) Maintenance of frequent patterns: a survey. In: Zhao Y, Zhang C, Cao L (eds) Post-mining of association rules: techniques for effective knowledge extraction, pp 273–293. Hershey, PA: information science reference. doi:10.4018/978-1-60566-404-0.ch014
17. Flake GW, Lawrence S, Giles CL, Coetzee FM (2002) Self-organization and identification of web communities. Computer 35(3):66–70
18. Franz M, Ward T, McCarley JS, Zhu W-J (2001) Unsupervised and supervised clustering for topic tracking. In: Paper presented at the proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval
19. Ganley D, Lampe C (2009) The ties that bind: social network principles in online communities. Decis Support Syst 47(3):266–274

20. Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826
21. Goyal A, Bonchi F, Lakshmanan LV (2008) Discovering leaders from community actions. In: Paper presented at the proceedings of the 17th ACM conference on information and knowledge management
22. Guimera R, Amaral LAN (2005) Functional cartography of complex metabolic networks. Nature 433(7028):895–900
23. Hofman JM, Wiggins CH (2008) Bayesian approach to network modularity. Phys Revi Lett 100(25):258701
24. Ito H, Iwama K (2009) Enumeration of isolated cliques and pseudo-cliques. ACM Trans Algorithms (TALG) 5(4):40
25. Ito H, Iwama K, Osumi T (2005) Linear-time enumeration of isolated cliques. In: Paper presented at the European symposium on algorithms
26. Kanawati R (2011) LICOD: Leaders identification for community detection in complex networks. In: Paper presented at the privacy, security, risk and trust (PASSAT) and 2011 IEEE third international conference on social computing (SocialCom)
27. Khorasgani RR, Chen J, Zaïane OR (2010) Top leaders community detection approach in information networks. In: Paper presented at the 4th SNA-KDD workshop on social network mining and analysis, Washington, DC
28. Kiss C, Bichler M (2008) Identification of influencers—measuring influence in customer networks. Decis Support Syst 46(1):233–253
29. Komusiewicz C, Hüffner F, Moser H, Niedermeier R (2009) Isolation concepts for efficiently enumerating dense subgraphs. Theor Comput Sci 410(38):3640–3654
30. Kumar R, Raghavan P, Rajagopalan S, Tomkins A (1999) Trawling the web for emerging cyber-communities. Comput Netw 31(11):1481–1493
31. Kuramochi M, Karypis G (2005) Finding frequent patterns in a large sparse graph. Data Min Knowl Discov 11(3):243–271
32. Lam HW, Wu C (2009) Finding influential ebay buyers for viral marketing a conceptual model of BuyerRank. In: Paper presented at the international conference on advanced information networking and applications
33. Lehmann S, Schwartz M, Hansen LK (2008) Biclique communities. Phys Rev E 78(1):016108
34. Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2008) Statistical properties of community structure in large social and information networks. In: Paper presented at the proceedings of the 17th international conference on world wide web
35. Lu D, Li Q, Liao SS (2012) A graph-based action network framework to identify prestigious members through member's prestige evolution. Decis Support Syst 53(1):44–54
36. Mislove AE (2009) Online social networks: measurement, analysis, and applications to distributed information systems. ProQuest, Rice University, Ann Arbor, United States
37. Nguyen NP, Dinh TN, Xuan Y, Thai MT (2011) Adaptive algorithms for detecting community structure in dynamic social networks. In: Paper presented at the Proceedings of the IEEE (INFOCOM 2011)
38. Nijssen S, Kok JN (2004) A quickstart in frequent structure mining can make a difference. In: Paper presented at the Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining
39. Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818
40. Pathak N, DeLong C, Banerjee A, Erickson K (2008) Social topic models for community extraction. In: Paper presented at the 2nd SNA-KDD workshop
41. Qi G-J, Aggarwal CC, Huang T (2012) Community detection with edge content in social media networks. Paper presented at the 2012 IEEE 28th international conference on data engineering
42. Sachan M, Contractor D, Faruquie TA, Subramaniam LV (2012) Using content and interactions for discovering communities in social networks. In: Paper presented at the proceedings of the 21st international conference on world wide web
43. Saito K, Yamada T, Kazama K (2008) Extracting communities from complex networks by the k-dense method. IEICE Trans Fundam Electron Commun Comput Sci 91(11):3304–3311
44. Satuluri V, Parthasarathy S (2009) Scalable graph clustering using stochastic flows: applications to community discovery. In: Paper presented at the proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining
45. Shen H, Cheng X, Cai K, Hu M-B (2009) Detect overlapping and hierarchical community structure in networks. Phys A Stat Mech Appl 388(8):1706–1712

46. Troussas C, Virvou M, Caro J, Espinosa KJ (2013) Mining relationships among user clusters in Facebook for language learning. In: Paper presented at the international conference on computer, information and telecommunication systems (CITS)
47. Uno T, Kiyomi M, Arimura H (2005) LCM ver. 3: collaboration of array, bitmap and prefix tree for frequent itemset mining. In: Paper presented at the proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations
48. Wasserman S, Faust K (1994) Social network analysis: methods and applications, vol 8. Cambridge University Press, Cambridge
49. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, San Francisco
50. Yan X, Han J (2002) gspan: graph-based substructure pattern mining. In: Paper presented at the Proceedings of the IEEE international conference on data mining (ICDM 2002)
51. Yang T, Jin R, Chi Y, Zhu S (2009) Combining link and content for community detection: a discriminative approach. In: Paper presented at the proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining
52. Zhou D, Manavoglu E, Li J, Giles CL, Zha H (2006) Probabilistic models for discovering e-communities. In: Paper presented at the proceedings of the 15th international conference on world wide web
53. Zhou Y, Cheng H, Yu JX (2009) Graph clustering based on structural/attribute similarities. Proc VLDB Endow 2(1):718–729
54. Zhu Z, Cao G, Zhu S, Ranjan S, Nucci A (2012) A social network based patching scheme for worm containment in cellular networks. In: Thai TM, Pardalos MP (eds) Handbook of optimization in complex networks: communication and social networks. Springer, Berlin, New York, pp 505–533
55. Zhuge H (2009) Communities and emerging semantics in semantic link network: discovery and learning. IEEE Trans Knowl Data Eng 21(6):785–799

Seyed Ahmad Moosavi received his B.S. degree in Computer Engineering from Institute of Motahari, Mashhad, Iran, in 2011 and M.S. Degree in Computer Software Engineering from Imam Reza International University, Mashhad, Iran, in 2014. He has been recognized as a first-rank student in Institute of Motahari, Mashhad. Also, he has been invited to study his degree (Master of Science) without any entrance exam as privilege for exceptional talents. His research interests are included data mining, especially data mining techniques in social networks and community detection and recommendation systems.

Mehrdad Jalali is currently an assistant professor in Computer Engineering Department, lead the Knowledge Engineering Lab (KEL Lab), and head of department at the Islamic Azad University of Mashhad. His research areas include data mining, machine learning, social networking, and semantic Web. He was staff researcher in Knowledge Technology Cluster, Artificial Intelligence Centre at MIMOS Technology Park Malaysia. Mehrdad Jalali received his bachelor degree from Islamic Azad University of Mashhad, Iran, in 1998 and Master degree in Artificial Intelligence from Science and Research University, Iran, in 2001 and his Ph.D. from Putra University of Malaysia in 2009. Mehrdad Jalali has published more than 80 papers in various journals, conferences, and book chapters. He is on the Editorial Boards of Journal of Information Technology on Engineering Design. He is active as reviewer for journals and international conferences.



Negin Misaghian is a Ph.D. student at Islamic Azad University, Sari Branch, Iran. She received her BS degree in Computer Software Engineering from Institute of Tabarestan, Chalus, Iran, in 2011 and obtained her M.S. degree in Computer Software Engineering from Imam Reza International University, Mashhad, Iran, in 2014. She has been recognized as first-rank student in Institute of Tabarestan and Imam Reza International University for her B.S. and M.S. degree, respectively. She has been invited to study both of her degree (Master of Science and Ph.D.) without any entrance exam as privilege for exceptional talents. Her research interests include data mining, recommendation systems, social networking. Another subject of her interest is requirements engineering as well.

Shahaboddin Shamshirband is a Senior Lecturer at the Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya. He is an Assistant Pofessor at the Department of Computer Science, Islamic Azad University, Chalous, Iran. His primary research area lies within computational intelligence, multi-agent systems, big data and machine learning in engineering applications of artificial intelligence. In addition, he is working on High Impact Research grant funded by University of Malaya. Currently, he is a Co-PI of the UMRG Programme by University of Malaya. He published more than 300 ISI-cited articles and numerous conference proceedings. He is a professional member of IEEE and also an editorial board member and reviewer for top journals (IEEE Transaction, Elsevier, and Springer).

Mohammad Hossein Anisi is a senior lecturer at the Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya. He obtained his Ph.D. from Universiti Teknologi Malaysia (UTM) while being awarded as the best postgraduate student. He worked as post-doctoral research fellow at UTM and was a member of pervasive computing research group (PCRG), a research group under K-Economy Research Alliance in Malaysia. His research interests lie in the area of wireless sensor networks and their applications, mobile ad hoc networks, and intelligent transportation systems. He has also collaborated actively with researchers in several other disciplines of computer science.